Empirical Bayes methods corrected for small numbers of tests

Marta Padilla and David R. Bickel *

February 27, 2012

Ottawa Institute of Systems Biology

Department of Biochemistry, Microbiology, and Immunology

University of Ottawa

451 Smyth Rd.

Ottawa, Ontario K1H 8M5

dbickel@uottawa.ca

Abstract

Histogram-based empirical Bayes methods developed for analyzing data for large numbers of genes, SNPs, or other biological features tend to have large biases when applied to data with a smaller number of features such as genes with expression measured conventionally, proteins, and metabolites. To analyze such small-scale and medium-scale data in an empirical Bayes framework, we introduce corrections of maximum

*The authors thank Ye Yang and Zhengmin Zhang for relevant discussions, Zhenyu Yang for proofreading, and Corey Yanofsky for both. We also thank the staff at Editage for copy editing the manuscript. This work was partially supported by the Faculty of Medicine of the University of Ottawa, by the Canada Foundation for Innovation, and by the Ministry of Research and Innovation of Ontario.

likelihood estimators (MLE) of the local false discovery rate (LFDR). In this context, the MLE estimates the LFDR, which is a posterior probability of null hypothesis truth, by estimating the prior distribution. The corrections lie in excluding each feature when estimating one or more parameters on which the prior depends. An application of the new estimators and previous estimators to protein abundance data illustrates how different estimators lead to very different conclusions about which proteins are affected by cancer.

The estimators are compared using simulated data of two different numbers of features, two different detectability levels, and all possible numbers of affected features. The simulations show that some of the corrected MLEs substantially reduce a negative bias of the MLE. (The best-performing corrected MLE was derived from the minimum description length principle.) However, even the corrected MLEs have strong negative biases when the proportion of features that are unaffected is greater than 90%. Therefore, since the number of affected features is unknown in the case of real data, we recommend an optimally weighted combination of the best of the corrected MLEs with a conservative estimator that has weaker parametric assumptions.

Keywords: empirical Bayes; local false discovery rate; medium-dimensional biology; medium-scale inference; minimum description length; penalized likelihood; reduced likelihood; selection bias; small-dimensional biology; small-scale inference; Type II maximum likelihood

1 Introduction

1.1 False discovery rates for genomics applications

In genomics, new technologies facilitate the simultaneous measurement of a wide variety of features, up to hundreds of thousands in number. Examples of such biological features include genes, locations in the brain, and single-nucleotide polymorphisms (SNPs) in genome-wide association studies. A multiple testing problem arises in the analysis of data involving N features $\langle X_1, X_2, \ldots, X_N \rangle$ of every individual belonging to two different groups, labeled *treatment* and *control* for convenience. For the ith feature and a corresponding effect size θ_i, a function T defines the statistic $T_i = T(X_i)$ that is used to test the null hypothesis that $\theta_i = \theta_0$, where θ_0 is the parameter value corresponding to no effect. For example, a common objective in genomics is to discover the genes that are differentially expressed between the treatment and control groups of individuals. Thus, gene expression data analysis involves testing N null hypotheses of equivalent expression.

Let A_i denote the variable indicating whether the ith alternative hypothesis is true. In the case of a two-sided alternative, $A_i = 1$ if $\theta_i \neq \theta_0$ but $A_i = 0$ if $\theta_i = \theta_0$. For example, $A_i = 1$ means the ith feature is affected by (or associated with) the treatment, disease, or other perturbation. The ith null hypothesis corresponds to a *discovery* of an effect if the statistic T_i falls within some *rejection region* $\mathcal{T}$, in which case, the ith null hypothesis is rejected. A discovery of an effect is a *false discovery* if there is no effect ($A_i = 0$); otherwise, it is a *true discovery* ($A_i = 1$).

The terminology follows Benjamini and Hochberg (1995), who introduced the *false discovery rate* (FDR) as an error measure for multiple testing. Many variants of the FDR can be found in literature, including the *Bayesian FDR* (Efron and Tibshirani, 2002) or *nonlocal FDR* (NFDR) (Bickel, 2011d) and the *local FDR* (LFDR) (Efron et al., 2001). In particular,

the NFDR is the probability that a null hypothesis is true, conditional on its rejection:

$$\Psi\left(\mathcal{T}\right)=\Pr\left(A_i=0|T_i\in\mathcal{T}\right)=\frac{E\left(N_0\left(\mathcal{T}\right)\right)}{E\left(N_+\left(\mathcal{T}\right)\right)},$$

where $N_0\left(\mathcal{T}\right)$ denotes the number of false discoveries and $N_+\left(\mathcal{T}\right)$ denotes the total number of discoveries (Efron, 2010). (Ψ is used to abbreviate $\psi\varepsilon\upsilon\delta\acute{\eta}\varsigma$, pseudo/false). The LFDR for the ith feature is defined as the probability that the null hypothesis is true given the statistic t_i, the observed realization of $T_i=T(X_i)$ (Efron, 2010). That is,

$$\psi_i=\Psi\left(\{t_i\}\right)=\Pr\left(A_i=0|T_i=t_i\right), \tag{1}$$

which assumes that T_i has a common probability density function g_{θ_0} conditional on the null hypothesis that $\theta_i=\theta_0$ and another probability density function g_{alt} conditional on the alternative hypothesis that $\theta_i\neq\theta_0$. According to Bayes's theorem,

$$\psi_i=P\left(\theta_i=\theta_0|t_i\right)=\frac{\pi_0 g_{\theta_0}\left(t_i\right)}{g\left(t_i\right)}, \tag{2}$$

where $\pi_0=P\left(\theta_i=\theta_0\right)$ is the expectation value of the proportion of null hypotheses that are true and $g\left(t_i\right)$ is the marginal probability density of the test statistic:

$$g\left(t_i\right)=\pi_0 g_{\theta_0}\left(t_i\right)+\left(1-\pi_0\right)g_{\text{alt}}\left(t_i\right). \tag{3}$$

As π_0 and $g\left(t_i\right)$ are unknown, they are estimated with empirical Bayesian methods to obtain the estimated LFDR by making substitutions into equations (2)-(3).

1.2 Motivation and overview

While high-dimensional biology involves measurements over numerous features, sometimes millions in number, small-dimensional biology involves measurements over fewer features.

Smaller-scale inference problems arise not only when the total data set represents a small number of genes, proteins, metabolites, voxels, or other features (e.g., Seifert et al., 2010), but also when there are subsets of a large number of features that have something in common that distinguishes them from the other features in the data set. For example, Efron (2008, §7) estimated the LFDR for each voxel as a member of a reference class of 82 voxels at the same physical location. The measurements of the other 15,461 voxels are less relevant to the truth of a null hypothesis corresponding to a voxel in the smaller reference class.

Unfortunately, the statistical methods that have been successfully applied to large-scale inference problems are not always directly applicable to inference problems involving considerably smaller dimensions. In particular, in the estimation of LFDR, commonly used methods of estimating the unknown parameters π_0 and $g(t_i)$ in equations (2) and (3) involve the histogram-based estimation of $g_{\text{alt}}(t_i)$ (e.g., Efron, 2004, 2007). While this is highly reliable for data sets with several thousand features (Yanofsky and Bickel, 2010; Montazeri et al., 2010), it has a high bias for data sets with small numbers of features. Therefore, special statistical methods are required when the number of features is too large for conventional hypothesis testing and yet too small for methods developed for an extremely large number of features. Hence, we propose new methods for the estimation of the LFDR in small-scale inference problems.

This paper is organized as follows. First, Section 2 recalls methods of eliminating a nuisance parameter by reducing the data vector x_i of the ith feature to a statistic $T(x_i)$ of smaller dimension. Section 3 reviews certain known LFDR estimators and presents the proposed LFDR estimation techniques. The application of the new LFDR estimators to a data set with 20 proteins is described in Section 4. The new LFDR estimators are then tested and compared using simulated data sets, as described in Section 5. Finally, Section 6 concludes the paper with a discussion. Asymptotic results are provided in Appendices A and B to explain the information-theoretic background behind one of the new estimators and to relate it to maximum likelihood estimation, respectively.

2 Data reduction and likelihood

Let $x \in \mathcal{X}$ be a vector of measurements of one feature. Note that since only one feature is considered in this section, the subscript *"i"* is not used, except in Example 3, where a generalization to N features is shown. The observed data vector $x \in \mathcal{X}$ is considered a realization of the random variable X of probability distribution $P_{\theta,\lambda}$ that admits a probability density function $f_{\theta,\lambda}$ with respect to some dominating measure, where $\theta \in \Theta$ is the parameter of interest and $\lambda \in \Lambda$ is the nuisance parameter. In the case of discrete X, the density function is defined with respect to the counting measure on $\mathcal{X}$. For some known $\theta_0 \in \Theta$, we have $\theta = \theta_0$ under the null hypothesis or narrow model and $\theta \neq \theta_0$ under the alternative hypothesis or wide model.

The following two types of likelihood correspond to different ways of reducing a vector x to a scalar statistic and of eliminating the nuisance parameter. Which of the two methods is appropriate depends on the original parametric family $\{f_{\theta,\lambda} : \theta \in \Theta, \lambda \in \Lambda\}$ and on which parameter is of interest.

2.1 Conditional likelihood

Consider the functions S and T such that $S(X)$ and $T(X)$ are statistics that together contain all the information in X. If $S(X)$ does not depend on θ and if the probability density function $g_\theta = f_\theta(\bullet | S(X) = S(x))$ of the data conditional on $S(x)$, the realized value of that statistic, does not depend on λ, then the function ℓ defined by

$$\ell(\theta) = g_\theta(T(x)) = f_\theta(T(x) | S(X) = S(x)) \tag{4}$$

is called the *conditional likelihood function* given $S(x)$. In analogy with equation (5), Severini (2000, §8.2.1) has

$$f_{\theta,\lambda}(x) = f_{\theta,\lambda}(S(x), T(x)) = g_\theta(T(x)) f_{\theta,\lambda}(S(x)),$$

where $f_{\theta,\lambda}$ can denote the probably density function of X, $\langle S(X), T(X)\rangle$, or $S(X)$, depending on the context.

Example 1. (Severini, 2000, Example 8.47). Suppose that X_1 is binomial $\langle n_1, \pi_1\rangle$, X_2 is binomial $\langle n_2, \pi_2\rangle$, and X_1 is independent of X_2. The parameter of interest is

$$\theta = \log \frac{\pi_1}{1-\pi_1} - \lambda,$$

where λ is the nuisance parameter

$$\lambda = \log \frac{\pi_2}{1-\pi_2}.$$

Then,

$$\log L(\theta, \lambda) = x_1\theta + S(x_1, x_2)\lambda - n_1 \log\left(1 + e^{\theta+\lambda}\right) - n_2 \log\left(1 + e^{\lambda}\right),$$

where $S(x_1, x_2) = x_1 + x_2 = s$ is sufficient. Then, taking $T(x_1, x_2) = x_1$, the conditional log-likelihood function given $S(x_1, x_2)$ is

$$\log \ell(\theta) = \log g_\theta(x_1) = \theta x_1 - \log K(\theta),$$

where

$$K(\theta) = \sum_{j=\max(0,s-n_2)}^{\min(n_1,s)} \binom{n_1}{j}\binom{n_2}{s-j} e^{j\theta}.$$

Conditional likelihoods are generally available whenever the parameter of interest is a natural parameter of an exponential family (Pawitan, 2001, §10.3). For details, see Severini (2000, §8.2.4). A recent application of the conditional likelihood function to genomics data can be found in Yang et al. (2011).

2.2 Marginal likelihood

Let T be a measurable function on $\mathcal{X}$. If, for each $\theta \in \Theta$, the probability density function g_θ of the *statistic* or *reduced data* $T(X)$ does not depend on the value of λ, then $\ell(\theta) = g_\theta(T(x))$ defines the *marginal likelihood function* ℓ.

If, in addition, the conditional distribution of X given $T(X) = T(x)$ does not depend on θ, then $T(X)$ is called *sufficient* for θ. In that case, no information about θ is lost in replacing X with $T(X)$:

$$\begin{aligned} f_{\theta,\lambda}(x) &= g_\theta(T(x))\, f_{\theta,\lambda}(x|T(X) = T(x)) \\ &= g_\theta(T(x))\, f_\lambda(x|T(X) = T(x)) \\ &= C g_\theta(T(x)), \end{aligned} \tag{5}$$

where C is constant in θ. The constant is unimportant because it drops out of likelihood ratios:

$$\frac{f_{\theta_1,\lambda}(x)}{f_{\theta_0,\lambda}(x)} = \frac{C g_{\theta_1}(T(x))}{C g_{\theta_0}(T(x))} = \frac{\ell(\theta_1)}{\ell(\theta_0)}$$

for any value of $\lambda \in \Lambda$.

Example 2. Suppose x and y are vectors of m and n values that realize the random variables X and Y of independent components drawn from normal distributions of unknown means ξ and η, respectively, and of a common unknown standard deviation σ. The parameter of interest is the inverse coefficient of variation defined by $\theta = (\xi - \eta)/\sigma$ with $\theta = 0$ as the null hypothesis and $\theta \neq 0$ as the alternative hypothesis; the parameter space here is $\Theta = \mathbb{R}^1$. A suitable statistic for data reduction is the two-sample t statistic

$$T(x, y) = \frac{\hat{\xi}(x) - \hat{\eta}(y)}{\hat{\sigma}(x, y)\sqrt{m^{-1} + n^{-1}}}, \tag{6}$$

where $\hat{\xi}$, $\hat{\eta}$, and $\hat{\sigma}^2$ are the usual unbiased estimators. Then $g_\theta(T(x, y))$, the probability

density of $T(X, Y)$ evaluated at the observation $\langle x, y \rangle$, is the noncentral Student t probability density with $m + n - 2$ degrees of freedom and noncentrality parameter $\left(m^{-1} + n^{-1}\right)^{-1/2} \theta$.

The next example encompasses data of multi-dimensional biology.

Example 3. Example 2 is extended to N genes, proteins, or other biological features such that $X_i \sim \mathrm{N}(\xi_i, \Sigma_{i,m})$ and $Y_i \sim \mathrm{N}(\eta_i, \Sigma_{i,n})$ correspond to the observed outcome $\langle x_i, y_i \rangle$ for the ith feature, where $i = 1, \ldots, N$ and $\Sigma_{i,k}$ is the diagonal covariance matrix of determinant σ_i^{2k}; thus, σ_i is the standard deviation of independent measurements of feature i. If whether or not there is an effect on feature i is much more important than the direction of that effect, the parameter of interest for feature i may be

$$\theta_i = |\xi_i - \eta_i| / \sigma_i, \tag{7}$$

the absolute value of the inverse coefficient of variation, with $\theta_i = 0$ as the null hypothesis, $\theta_i > 0$ as the alternative hypothesis, and $\Theta = [0, \infty)$ as the parameter space. Then $T(x_i, y_i)$ is the absolute value of the two-sample t statistic for $\langle x_i, y_i \rangle$ according to equation (6), and $T(X_i, Y_i)$ is distributed as the absolute value of a variate from the noncentral Student t distribution with $m+n-2$ degrees of freedom and noncentrality parameter $\delta_i = \left(m^{-1} + n^{-1}\right)^{-1/2} \theta_i$. Thus, the density $g_{\theta_i}(T(x_i, y_i))$ for the ith feature is the probability density of $T(X_i, Y_i)$ evaluated at $\langle x_i, y_i \rangle$. Bickel (2011b,e) illustrated different methods of penalized maximum likelihood estimation of the LFDR under this model.

Severini (2000, §8.3) and Schweder and Hjort (2002) provide additional examples of the marginal likelihood, also called the *reduced likelihood* and not to be confused with the likelihood integrated with respect to a prior distribution.

3 Local false discovery rate estimation

As mentioned in Section 1, previous estimators of FDR and LFDR are highly biased for a moderate or small number of hypotheses. We present several strategies in this section to reduce that bias.

3.1 Previous LFDR estimators

In this subsection, we review the previous LFDR estimators that lay the foundations on which our new estimators are constructed.

3.1.1 LFDR estimates based on other false discovery rates

Recall from Section 1 that the ith null hypothesis is rejected if the statistic t_i falls within some *rejection region* $\mathcal{T}$. To avoid the specification of such a rejection region $\mathcal{T}$, an estimated q-value $q(p_i)$ is commonly calculated for the ith p-value p_i among the N p-values. The rejection region $\mathcal{T}_\alpha$ is a function of the significance level α, the usual Type I error rate of rejecting the ith null hypothesis if and only if $p_i \leq \alpha$; thus, the estimated q-values, herein called *q-values* to follow contemporary terminology (Hong et al., 2009), are given by

$$q(p_i) = \min_{\alpha \in [p_i, 1]} \widehat{\mathrm{pFDR}}(\mathcal{T}_\alpha)\,;\, [i = 1, \ldots, N]\,, \tag{8}$$

where $\widehat{\mathrm{pFDR}}(\mathcal{T}_{p_i})$ is an estimate of the positive FDR (pFDR) on the rejection region $\mathcal{T}_{p_i}$ (Storey, 2002). Thus, the q-value is the lowest estimated pFDR at which the ith null hypothesis is rejected. Because the latter effectively uses 1 as an estimate of π_0, it will be called QV1 in order to distinguish it from $q(p_i)$, which is called QV.

In addition, conservative LFDR estimators based on the binomial distributions have been proposed by Bickel (2011d). The estimator that Bickel (2011d) called the "MLE" is renamed in this paper to avoid confusion with the estimator addressed in the next subsection. We denote the version that uses the estimate of π_0 described in Storey (2002) as the *binomial-*

based estimator (BBE) to distinguish it from BBE1, which instead uses 1 as an estimate of π_0.

3.1.2 Maximum likelihood estimator

The maximum likelihood estimator (MLE) described in this subsection will be called the *leave-zero-out* (L0O) method for reasons given in Section 6.1. The LFDR is estimated under the assumption that both the null-hypothesis density function g_{θ_0} and the alternative-hypothesis density function g_{alt} of equations (2)-(3) are members of $\{g_\theta : \theta \in \Theta\}$, a parametric family of probability density functions indexed by the interest parameter θ, which is a member of some parameter space Θ. Thus, $g_{\text{alt}} = g_{\theta_{\text{alt}}}$, where $\theta_{\text{alt}} \in \Theta$ is unknown and not equal to the known $\theta_0 \in \Theta$. Any nuisance parameter must be eliminated, perhaps by using one of the two methods explained in Section 2.

For the ith feature, the data vector x_i is reduced to a scalar statistic t_i, as in Examples 1-3. Therefore, $g_{\theta_0}(t_i)$ and $g_{\theta_{\text{alt}}}(t_i)$ denote the probability densities for the reduced data under the null hypothesis and the alternative hypothesis, respectively. The true value of the LFDR for the ith feature is, according to equations (2)-(3) and $g_{\text{alt}} = g_{\theta_{\text{alt}}}$,

$$\psi_i = \frac{\pi_0 g_{\theta_0}(t_i)}{\pi_0 g_{\theta_0}(t_i) + (1-\pi_0) g_{\theta_{\text{alt}}}(t_i)}, \tag{9}$$

which is unknown since θ_{alt} and π_0 are unknown.

The L0O method involves the estimation of the parameters π_0 and θ_{alt}. These estimated parameters $\left\langle \hat{\theta}^{\text{L0O}}, \hat{\pi}_0^{\text{L0O}} \right\rangle$ are the maximum likelihood estimates of the true parameters given by

$$\left\langle \hat{\theta}^{\text{L0O}}, \hat{\pi}_0^{\text{L0O}} \right\rangle = \arg \sup_{\langle \theta, \pi_0 \rangle \in \Theta \times [0,1]} \prod_{j=1}^{N} \left(\pi_0 g_{\theta_0}(t_i) + (1-\pi_0) g_\theta(t_i) \right). \tag{10}$$

Therefore, with substitution into equation (2), the estimated LFDR for the ith feature is

$$\hat{\psi}_i^{\text{L0O}} = \frac{\hat{\pi}_0^{\text{L0O}} g_{\theta_0}(t_i)}{\hat{\pi}_0^{\text{L0O}} g_{\theta_0}(t_i) + (1-\hat{\pi}_0^{\text{L0O}}) g_{\hat{\theta}^{\text{L0O}}}(t_i)}. \tag{11}$$

This estimator has been used with marginal likelihood (Yang and Bickel, 2010; Bickel, 2011e) and conditional likelihood (Yang et al., 2011). Similarly, Muralidharan (2010) had estimated the LFDR by maximizing the likelihood over exponential families.

3.2 New LFDR estimators

Here, 5 novel LFDR estimators are proposed: 3 are corrected MLEs, and the other 2 are related to the BBE. The corrected MLEs are based on equation (2). The fourth technique is an approximation of the BBE, and the last new estimator is a combination of the BBE and one of the corrected MLEs.

3.2.1 Corrected MLEs

The three methods presented here correct the bias of the L0O method that results from using the same statistic t_i to evaluate the density functions and to estimate π_0 and θ_{alt}. This is accomplished by removing dependence of the estimators on t_i prior to evaluating the density functions at t_i. While that negative bias vanishes as the number of features increases (Appendix B), it can be unacceptably large for small numbers of features.

The first corrected MLE is called the *minimum description length* (MDL) method. Although the method was inspired by the MDL principle (Appendix A), the general idea of estimating a prior on the basis of exchangeable features other than the feature under consideration is implicit in Goodman (2004); cf. Gastpar et al. (2010) and J. Cuzick's discussion of Aitkin (1991). The MDL method uses modified estimates of parameters π_0 and θ_{alt} for the ith feature, denoted as $\left\langle \hat{\theta}_i^{\text{MDL}}, \hat{\pi}_{0i}^{\text{MDL}} \right\rangle$ for $i \in \{1, \ldots, N\}$. These estimated parameters are obtained by maximizing the likelihood function:

$$\left\langle \hat{\theta}_i^{\text{MDL}}, \hat{\pi}_{0i}^{\text{MDL}} \right\rangle = \arg \sup_{\langle \theta, \pi_0 \rangle \in \Theta \times [0,1]} \prod_{j=1,\, j \neq i}^{N} \left(\pi_0 g_{\theta_0}(t_i) + (1 - \pi_0)\, g_\theta(t_i) \right). \tag{12}$$

Note that the product is obtained over all features except for the ith feature. Accordingly,

the MDL estimator of LFDR for the ith feature is given by

$$\hat{\psi}_i^{\mathrm{MDL}} = \frac{\hat{\pi}_{0i}^{\mathrm{MDL}} g_{\theta_0}}{\hat{\pi}_{0i}^{\mathrm{MDL}} g_{\theta_0}(t_i) + (1 - \hat{\pi}_{0i}^{\mathrm{MDL}})\, g_{\hat{\theta}_{\mathrm{i}}^{\mathrm{MDL}}}(t_i)}. \tag{13}$$

The second corrected MLE estimator, called *leave-one-out* (L1O), is the same as the MDL except that the L0O estimate of π_0 is used instead of $\hat{\pi}_{0i}^{\mathrm{MDL}}$. Therefore, in L1O, three steps are involved. First, the parameters $\left\langle \hat{\theta}^{\mathrm{L0O}}, \hat{\pi}_0^{\mathrm{L0O}} \right\rangle$ are calculated from the likelihood function (10) involved in the L0O method, which includes all the features. Second, similar to MDL, the likelihood function involving all features, except for the ith feature, is maximized for every feature using the $\hat{\pi}_0^{\mathrm{L0O}}$ obtained in the previous step. Therefore, in this step, the interest parameter for all $i \in \{1, \ldots, N\}$ is estimated as

$$\hat{\theta}_i^{\mathrm{L1O}} = \arg\sup_{\theta \in \Theta} \prod_{j=1,\, j \neq i}^{N} \left(\hat{\pi}_0^{\mathrm{L0O}} g_{\theta_0}(t_j) + \left(1 - \hat{\pi}_0^{\mathrm{L0O}}\right) g_{\theta}(t_j) \right), \tag{14}$$

leading to the L1O estimator of LFDR for the ith feature:

$$\hat{\psi}_i^{\mathrm{L1O}} = \frac{\hat{\pi}_0^{\mathrm{L0O}} g_{\theta_0}}{\hat{\pi}_0^{\mathrm{L0O}} g_{\theta_0}(t_i) + (1 - \hat{\pi}_0^{\mathrm{L0O}})\, g_{\hat{\theta}_{\mathrm{i}}^{\mathrm{L1O}}}(t_i)}. \tag{15}$$

The MDL and L1O strategies eliminate bias from a double use of feature data. However, when there is only a single affected feature, the MDL and L1O do not use any information about θ_{alt} to estimate the LFDR of the only affected feature, introducing considerable bias in estimating θ_{alt}.

To overcome this defect, we introduce the third corrected MLE, called the *leave-half-out* (L$^1/_2$O) estimator. Like L1O, L$^1/_2$O includes information about the ith feature through $\hat{\pi}_0^{\mathrm{L0O}}$; furthermore, half of the information about each left-out feature is also included in the L$^1/_2$O through the likelihood function. Such a function is a *weighted likelihood function*, where the contribution of the ith feature to the overall likelihood function is corrected by a *weight* w_{ij}

given by

$$w_{ii}(\nu) = \frac{\nu}{\nu + (N-1)};\; w_{ij}(\nu) = \frac{1}{\nu + (N-1)}\;[j \neq i],$$

where $\nu \in [0,1]$ is the information (log-likelihood) weight of t_i relative to each t_j for the purpose of estimating the parameter of interest. Thus, the weights satisfy $\sum_{j=1}^{N} w_{ij} = 1$. The ν-weighted likelihood function for feature i is

$$L_i(\pi_0,\, \theta_{\text{alt}};\, t_i, \nu) = \prod_{j=1}^{N} \left(\pi_0 g_{\theta_0}(t_j) + (1-\pi_0)\, g_{\theta_{\text{alt}}}(t_j)\right)^{w_{ij}(\nu)}, \tag{16}$$

and the maximum ν-weighted likelihood is

$$\hat{\theta}_i^{\text{L}\nu\text{O}} = \arg\sup_{\theta \in \Theta} L_i(\hat{\pi}_0^{\text{L0O}},\, \theta;\, t_i, \nu), \tag{17}$$

degenerating to the L0O and L1O estimators when $\nu = 1$ and $\nu = 0$, respectively. Thus, $\hat{\theta}_i^{\text{L}\nu\text{O}}$ may be considered the *leave-ν-out* (LνO) estimator.

The proposed L1/2O method includes exactly half of the information about the ith feature in its likelihood function by setting $\nu = 1/2$. Therefore, the new estimator $\hat{\theta}_i^{\text{L1/2O}}$ for all $i \in \{1, \ldots, N\}$ is given by the maximization of the weighted likelihood function according to equations (16)-(17). With such estimated parameters $\left\langle \hat{\theta}_i^{\text{L1/2O}}, \hat{\pi}_0^{\text{L0O}} \right\rangle$, the LFDR for the ith feature ($\hat{\psi}_i^{\text{L1/2O}}$) can be estimated with equation (15), replacing $\hat{\theta}_i^{\text{L1O}}$ with $\hat{\theta}_i^{\text{L1/2O}}$, and analogously for $\hat{\psi}_i^{\text{L}\nu\text{O}}$ given any ν between 0 and 1.

Weighted likelihoods have been reviewed by Hu and Zidek (2002) and applied to the quantification of evidence by Bickel (2011b).

3.2.2 BBE-related LFDR estimators

A method for approximating the BBE (Bickel, 2011d) is also presented here. BBE attempts to estimate the LFDR more conservatively than q-values, which were not originally designed for LFDR estimation. In this section, we denote the q-values as q, which refers to either QV

or QV1 (see Section 3), and ρ_i denotes the rank of the q-values corresponding to the ith feature, such that $q_{(1)} \leq q_{(2)} \leq ... \leq q_{(N)}$. The new proposed method directly assigns twice the rank of the q-value $q_{(2\rho_i)}$ to the LFDR estimate of the ith feature with the corresponding q-value $q_{(\rho_i)}$. Therefore, we define *(estimated) r-values* as

$$r(q_i) = \begin{cases} q_{(2\rho_i)} & \text{if } \rho_i \leq N/2 \\ 1 & \text{if } \rho_i > N/2. \end{cases} \tag{18}$$

We employ analogous notation for r-values, i.e., RV when it uses QV and RV1 when it uses QV1. Our aim is to verify that RV and RV1 approximate BBE and BBE1, respectively.

Finally, for reasons given in Section 6.1, we combine BBE and MDL into an estimator that leverages the strengths of each. Specifically, the *MDL-BBE* is the linearly combination of the other two estimators with weights that are optimal for the hedging game of Bickel (2011c).

4 Application

In Alex Miron's laboratory at the Dana-Farber Cancer Institute, the abundance levels of 20 plasma proteins of 55 women with HER2-positive breast cancer, 35 women with ER/PR-positive breast cancer, and 64 healthy women (Li, 2009) were measured. The respective data vectors $x_1^{\text{HER2}}, \ldots, x_{20}^{\text{HER2}}$, $x_1^{\text{ER/PR}}, \ldots, x_{20}^{\text{ER/PR}}$, $y_1, \ldots, y_{20}$ were created by adding the first quartile of the abundance levels (over the 64 healthy women and over all proteins) to each abundance level and by taking natural logarithms of the resulting sums; similar conservative prepossessing steps have worked well with gene expression data (Bickel, 2002).

The preprocessed data were modeled as normally distributed, as illustrated in Example 3. Following the notation of the example, ξ_i^{HER2}, $\xi_i^{\text{ER/PR}}$, and η_i are the expectation values of X_i^{HER2}, $X_i^{\text{ER/PR}}$, and Y_i, respectively, and are as such interpretable as population levels of the abundance of protein i. The parameters of interest are $\theta_i^{\text{HER2}} = \left|\xi_i^{\text{HER2}} - \eta_i\right| / \sigma_i$ and

$\theta_i^{\text{ER/PR}} = \left|\xi_i^{\text{ER/PR}} - \eta_i\right| / \sigma_i$, the standardized levels of the ith protein's abundance relative to the healthy controls. In this context, the LFDR of each protein is a posterior probability that its average abundance level is not affected by cancer.

The data were analyzed according to the distributions of $T\left(X_i^{\text{HER2}}, Y_i\right)$ and $T\left(X_i^{\text{ER/PR}}, Y_i\right)$ given in Example 3. The methods of estimating the LFDR described in Section 3.1 were applied to the proteomics data, namely, MDL, L0O, L1O, L1/2O, BBE, BBE1, RV, RV1, and MDL-BBE. The results are shown in Figures 1 and 2, which represent LFDR versus the estimated protein abundance ratio and p-value, respectively. All figures show results for the HER2-positive and ER/PR-positive groups separately. The volcano plot (Figure 1) indicates that the proteins most affected by cancer, showing estimated abundance ratios furthest from unity, have LFDR estimates close to zero, while higher LFDR estimates correspond to proteins with estimated abundance ratios close to unity. From the results shown in both figures, we can see that the selection of the LFDR estimator is crucial because for thresholds of the estimated LFDR between 0 and 0.2, many proteins would be considered affected or unaffected by cancer, depending on the method. BBE1 and RV1 were omitted from the figures to ensure legibility.

5 Simulations

In this section, the performance of the LFDR estimators described in Section 3.1 is compared using simulated protein abundance data. Such methods are MDL, L0O, L1O, L1/2O, BBE, BBE1, RV, RV1, and a combination of MDL and BBE. The design of each data set is patterned after that of Sections 3.1.2 and 4. It consists of abundance levels of N proteins for two groups, *sick* and *healthy,* each containing 5 individuals, for total of 10 abundance levels per protein. For the ith protein, the log-abundance data are drawn from a normal distribution with variance $\sigma^2 = 1$ and mean equal to 0, except for the proteins affected by the disease, which have mean $\xi_{\text{alt}} > 0$ in the sick group. To represent both barely detectable

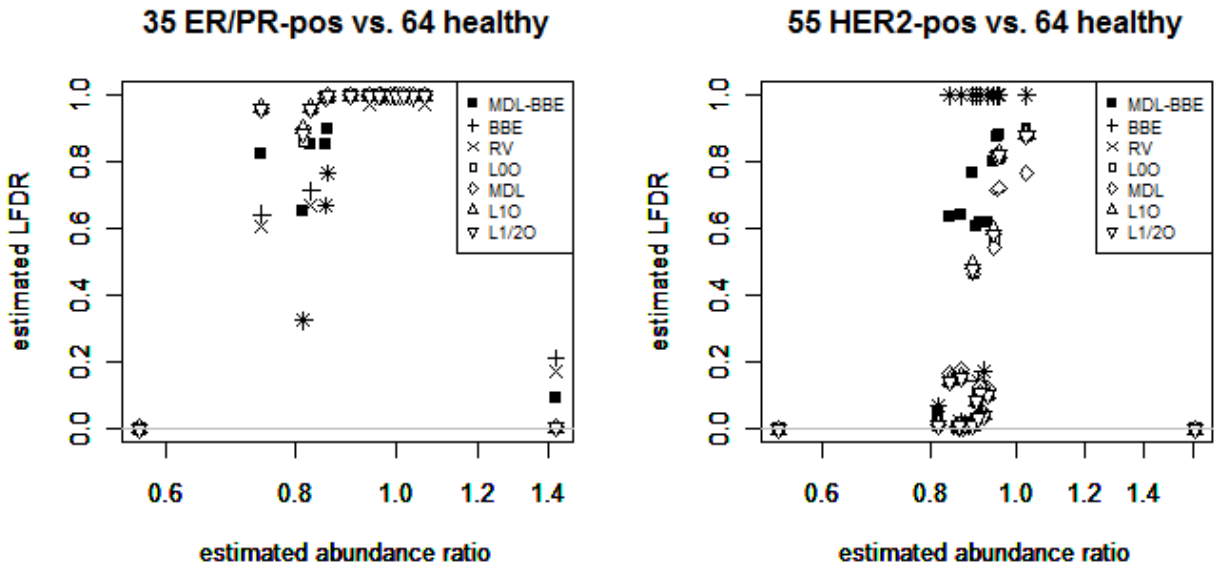

Figure 1: Volcano plot representing LFDR for protein abundance of both groups, HER2-positive and ER/PR-positive women, relative to healthy women, estimated by using different LFDR estimators and represented versus the estimated protein abundance ratio. The LFDR estimators are MDL, L0O, L1O, L$^1/_2$O, BBE, RV, and MDL-BBE.

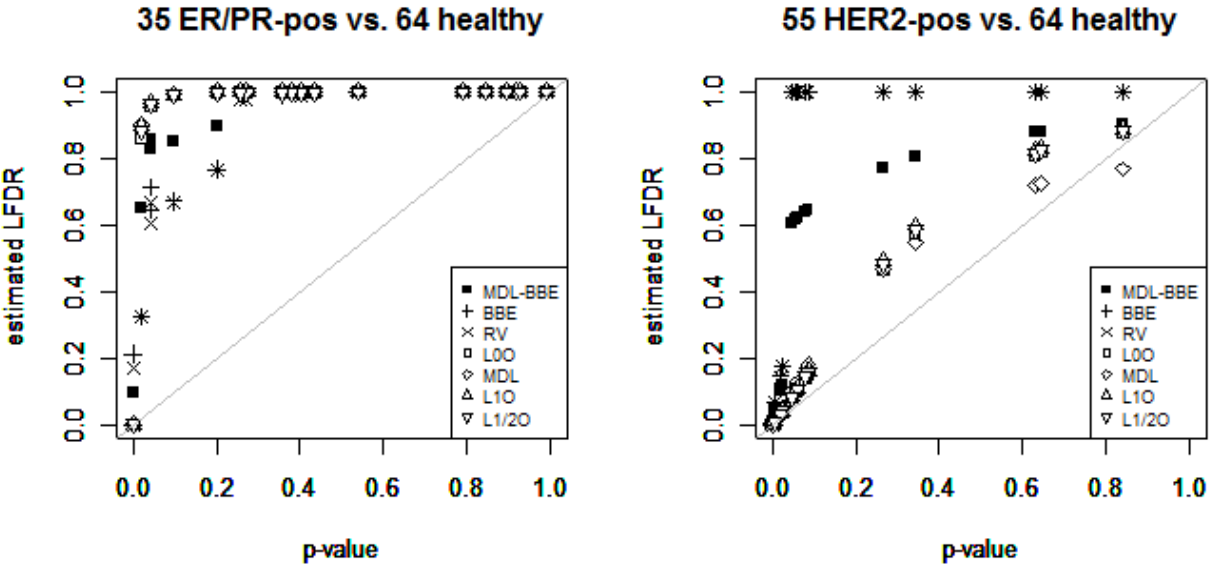

Figure 2: LFDR for protein abundance of both groups of women with breast cancer, HER2-positive and ER/PR-positive, relative to healthy women, estimated by using different estimators and represented versus p-value. The LFDR estimators are MDL, L0O, L1O, L$^1/_2$O, BBE, RV, and MDL-BBE.

and highly detectable differences between the null and alternative distributions, we consider two values for the effect size, a low value ($\xi_{\text{alt}} = 1.5$) and a high value ($\xi_{\text{alt}} = 4$) relative to the standard deviation ($\sigma = 1$). Therefore, we have two values for the positive noncentrality parameter $\delta_{\text{alt}} = (m^{-1} + n^{-1})^{-1/2}\, \theta_{\text{alt}}$, where, in agreement with equation (7),

$$\theta_{\text{alt}} = |\xi_{\text{alt}} - 0|/\sigma = \xi_{\text{alt}}, \tag{19}$$

and m and n are the numbers of individuals in the sick and healthy group, respectively ($m = n = 5$). Therefore, the distribution of the affected proteins in the sick group has $\delta_{\text{alt}} = 2.4$ in the low-effect simulations and $\delta_{\text{alt}} = 6.3$ in the high-effect simulations. By contrast, the noncentrality parameter values are 0 for all the unaffected proteins and for all the proteins of the healthy group. Then, the LFDR estimators are compared with regard to the number of proteins in each data set and the number of affected proteins for 20 simulated data sets of each configuration.

To facilitate the comparison among the different LFDR estimators and for specific verification of the similarities between BBE and RV and BBE1 and RV1, we estimated each estimator's *bias*, the mean (over all proteins) of the expectation value of the difference between the estimate and true LFDR. For each LFDR estimator, that bias is estimated by the mean difference between the estimated LFDR and the true LFDR, where the mean is over the simulations as well as the proteins. Thus, 60 LFDR estimates are averaged when the data set has 3 proteins (mean over 3 proteins and over 20 simulations) and 300 LFDR values when the data set has 15 proteins (mean over 15 proteins and over 20 simulations). The true value is calculated using equation (9) with the proportion of proteins that are unaffected as π_0 and with the value of θ_{alt} given by equation (19).

The results are shown in Figure 3, where the estimated bias of the LFDR is represented as a function of the number of affected proteins, for each number of proteins in the data set and for two different levels of detectability. Although we studied the behavior of the methods

separately for affected and unaffected proteins, the figures show the estimated biases of the LFDR averaged over all proteins. Figure 3, plots (a) and (b), show the results for a data set with 3 and 15 proteins, respectively, and for the high level of detectability. Plots (c) and (d) are the same except that they correspond to the low detectability level. For better legibility of the figures, RV1 and BBE1 are not displayed because they have excessively high estimated bias averaged over either affected or unaffected proteins or averaged over all proteins.

It can be seen from Figure 3 that the LFDR estimates depend on the number of proteins, the number of affected proteins, and the detectability level. Note that in the plots, the contribution of the bias from the affected proteins increases as the number of affected proteins increases because the protein-averaged results are, in effect, weighted according to the number of affected or unaffected proteins. The estimators BBE1 and RV1 are not displayed because the values of their (positive) biases are much higher than those of the other estimators. However, the biases of RV and BBE are more moderate, especially when few proteins are affected.

6 Discussion

6.1 Evaluation of the LFDR estimators

Some differences in estimator performance depend on the value of δ_{alt}, the noncentrality parameter. L0O and L$^1/_2$O work very well when δ_{alt} is high, regardless of the number of features in the data set (Figure 3, (a)-(b)) and when there is at least one affected feature. When there is no affected feature, both estimators have highly negative bias (about -0.25). When δ_{alt} is high, MDL and L1O perform similarly to L0O and L$^1/_2$O, except when there is only one affected feature in the data set, in which case MDL and L1O have excessively high positive biases for the affected feature. Those biases are not seen in the plots since they are averaged over all features. These biases result from the fact that MDL and L1O do not use the data of the given feature to estimate δ_{alt}. Thus, MDL and L1O cannot effectively

estimate δ_{alt} when only one feature is affected, which results in such a noticeable high positive bias when δ_{alt} is high. L1/2O overcomes that drawback by including half the information on the unique affected feature in its likelihood function (16). In contrast, BBE and RV are less biased than the other estimators when no features are affected. However, the values of their conservative (positive) biases increase with the number of affected features. On the other hand, when δ_{alt} is low (Figure 3, (c)-(d)), all the corrected MLEs are negatively biased when there is no affected feature, and BBE and RV have positive biases.

In addition, by comparing the four plots in Figure 3, we can see that BBE and RV are extremely similar; only slight differences appear in cases of few affected proteins. Moreover, the methods gave similar estimates in the application to real protein data (Figures 2 and 1).[1]

Therefore, since BBE-related estimators show a small bias for no or a few affected features and since corrected MLEs perform better when most of the features of the data set are affected, we consider a new LFDR estimator as the weighted combination of representative estimators of each type (corrected MLEs and BBE-related estimators): the MDL and the BBE. Based on performance with 3 features and low δ_{alt}, we choose to combine MDL and BBE because MDL has the lowest absolute value of the bias among the corrected MLEs and because the BBE is simpler than the RV but is similar in performance. Then we applied the same MDL-BBE method to all cases. The MDL-BBE is an optimal linear combination of the MDL and the BBE (Section 3.2.2).

To summarize the findings for each method and each total number of features in the data set, Table 1 reports the most extreme values and the median of the biases for $\pi_0 \geq 90\%$ over the numbers of affected features and over both values of δ_{alt}. We can see from this table that these values are very similar among the methods of the same type. Corrected MLEs have the most negative biases, and BBE-related methods have the highest positive biases. As Table 1 indicates, the MDL-BBE succeeds in substantially reducing the negativity of the

[1]However, we found in unpublished work that these estimators diverge more for an application to a large-scale proteomics data set.

LFDR	**all π_0**		**$\pi_0 \geq$90%**	
Estimators	**3 features**	**15 features**	**3 features**	**15 features**
MDL-BBE	**0.13 [−0.10, 0.41]**	**0.01 [−0.20, 0.31]**	**−0.1**	**−0.13 [−0.2, 0.01]**
BBE	**0.19 [−0.07, 0.69]**	**0.01 [−0.11, 0.55]**	**−0.07**	**−0.07 [−0.11,−0.01]**
RV	0.18 [−0.08, 0.69]	0.00 [−0.13, 0.55]	−0.08	−0.09 [−0.13,−0.02]
MDL	**0.02 [−0.13, 0.12]**	**0.00 [−0.30, 0.07]**	**−0.13**	**−0.19 [−0.30, 0.02]**
L0O	−0.02 [−0.22, 0.16]	−0.01 [−0.34, 0.08]	−0.18	−0.17 [−0.26,−0.01]
L1O	0.02 [−0.17, 0.20]	0.00 [−0.30, 0.08]	−0.17	−0.16 [−0.24, 0.03]
L1/2O	−0.02 [−0.22, 0.18]	0.00 [−0.32, 0.08]	−0.17	−0.17 [−0.24,−0.01]

Table 1: Median [minimum, maximum] values of the biases of all the LFDR estimators over all π_0 and over all $\pi_0 \geq 90$ %, over the numbers of affected features, and over both values of the noncentrality parameter. Separate values are given for each total number of features in the data set.

worst-case bias of the MDL and substantially reducing the highly conservative worst-case bias of the BBE. In short, the MDL-BBE does not suffer from the main drawbacks of the other estimators.

Since the focus on the worst-case performance can lead to an overly pessimistic assessment of small-scale estimation of the LFDR, the median values are also reported in Table 1. They indicate that while estimation is somewhat unreliable for some estimators when there are only 3 features, it is reliable for all estimators when there are 15 features. Even so, the reported absolute values of the biases should be regarded as lower bounds since they were computed under the independence of features. Further, since the simulations use the same family of distributions as the MLE-related estimators, they perform better in the simulations that they would with real data.

6.2 Conclusions

In this paper, we proposed several LFDR estimators to give reliable results for small-scale inference. We compared them on simulated data sets and illustrated their use on a protein-abundance data set that illustrates that different conclusions would be drawn on the basis of different estimators. The performance of such methods depends on the number of features, number of affected features, and values of the unknown parameters. Simulations showed that

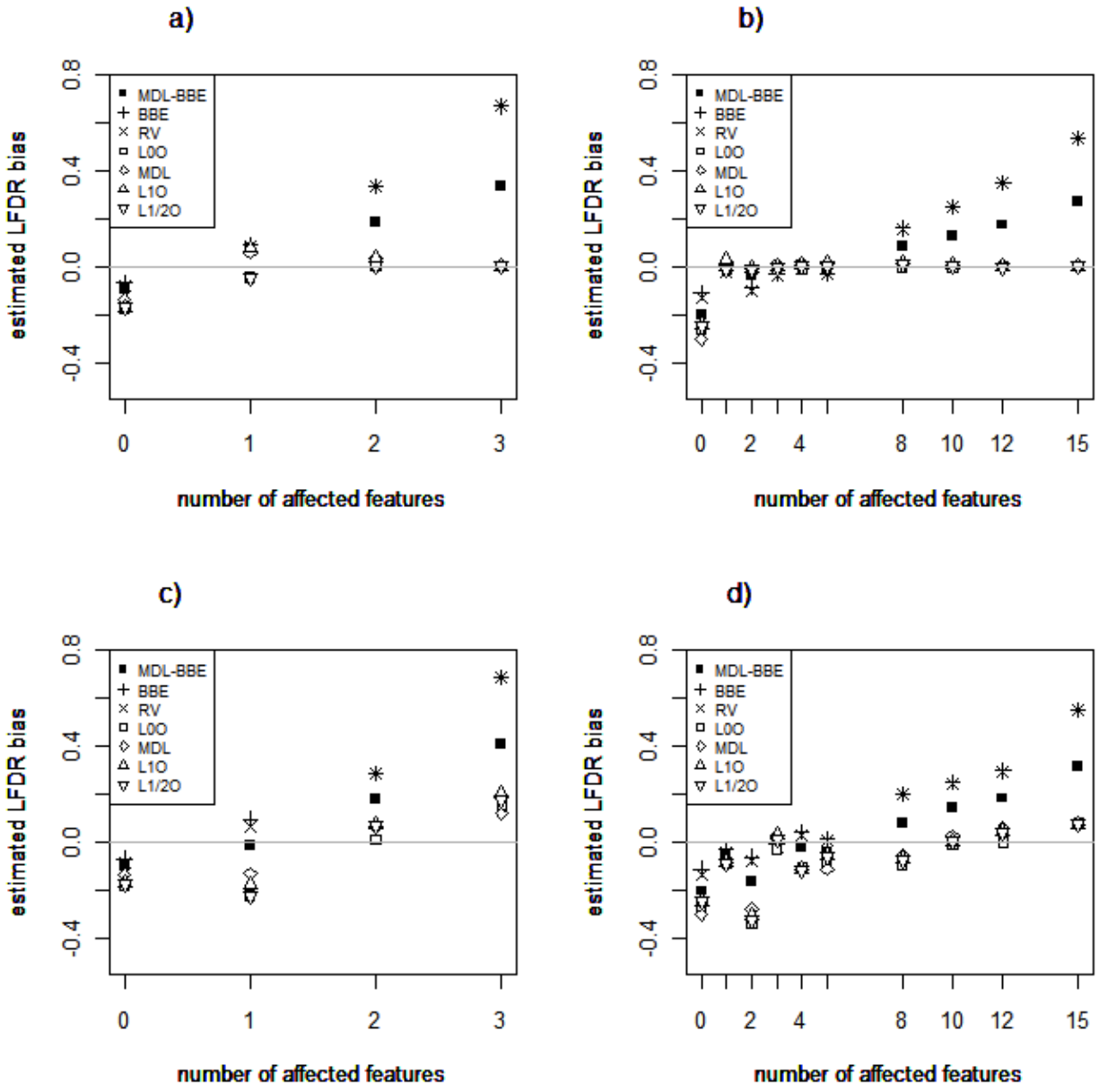

Figure 3: Estimated bias of several LFDR estimators for an artificial data set with 3 features ((a) and (c)) and 15 features ((b) and (d)) and cases of high $\delta_{\text{alt}} = 6.3$ ((a) and (b)) and low noncentrality parameter $\delta_{\text{alt}} = 2.4$ ((c) and (d)) versus increasing number of affected features. The LFDR estimators are MDL, L0O, L1O, L$^1/_2$O, BBE, RV, and a combination of MDL and BBE.

the corrected MLEs have very low biases in all cases when more than 50% of the features in the data set are affected, even for a data set with only 3 features. However, when the proportion of affected features is very small, these methods have excessively negative biases. In contrast, BBE and RV have excessively large biases when there is a high proportion of affected features. Furthermore, this bias increases as the number of affected features increases in the data set. Therefore, the weighted combination of an adjusted-MLE (MDL) and a conservative method (RV or BBE) may represent the safest solution for a general scenario in which the number of affected features is unknown.

Colophon

We used the following packages of R (R Development Core Team, 2008): `Biobase` (Gentleman et al., 2004) and `qvalue` (Dabney et al., 2011) from Bioconductor; `locfdr` (Efron, 2007), `fBasics` (Wuertz, 2010), and `distr` (Ruckdeschel et al., 2006) from the CRAN repository.

Appendix A: Methods motivating the new MDL method

This appendix uses the MDL principle to explain the statistical methods that led to the MDL method defined by equation (13). This appendix also has results that lay the foundation for the operating characteristics of the estimator given in Appendix B. A simple explanation of basic MDL-theoretic ideas in terms of hypothesis testing is available in the appendix of Bickel (2011b). See Rissanen (2007), Barron et al. (1998), Grünwald (2007), and Bickel (2011a) for other introductions to the MDL principle of model selection.

Since θ_{alt} is unknown, it will be replaced with a parameter value chosen to minimize the codelength of the data according to MDL theory, in which the length of a codeword is the number of independently selected binary digits of equal probability that achieve the joint probability of that codeword (Rissanen, 2007). The availability of measurements pertaining to features other than the inference target enables the construction of a universal codelength

function and a close approximation that is computationally more convenient. The idea that statistical inference minimizes universal codelength functions is called the *MDL principle* and is often formalized in terms of minimax problems.

Minimum description length concepts

The theory of this section is presented in terms of a parametric family that is free from nuisance parameters. In many cases, such a family can be derived using one of the data reduction methods of Section 2.

Under the MDL framework, each scheme † for coding the data under the alternative hypothesis corresponds to a codelength function $L^\dagger$ on $\mathcal{X}$ and thus to a *compressing probability density function* $g^\dagger$ selected from the parametric family $\{g_\theta : \theta \in \Theta\}$ before observing $T(x)$, the realized value of the statistic, with the goal of minimizing the codelength $L^\dagger(T(x)) = -\log g^\dagger(T(x))$. Since θ_0 is known, the probability density function of the statistic under the null hypothesis is known to be g_{θ_0}, which compresses the data with respect to the null model. Accordingly, the codelength function L^0 relative to the null hypothesis is that specified by $L^0(T(x)) = -\log g_{\theta_0}(T(x))$. Since the base of the logarithm is arbitrary, the inverse logarithm is denoted by $\log^{-1}\bullet$ rather than by $\exp\bullet$ or by $2^\bullet$.

Suppose, as in Example 3, that there is a vector x_i of measurements for each of the N features and that the data are reduced to the statistics $T(x_1), \ldots, T(x_N)$. With $L_i^\dagger(T(x_i))$ as the codelength of $T(x_i)$ relative to the alternative hypothesis, $\Delta_i^\dagger(T(x_i)) = L_i^\dagger(T(x_i)) - L^0(T(x_i))$ is the *information in* $T(x_i)$ *for discrimination* favoring the null hypothesis over the alternative hypothesis; cf. Bickel (2011b,a). A difference in null and alternative codelengths has been called a "universal test statistic" (Rissanen, 1987); however, that term can cause confusion with $T(X_i)$.

Example 4. If the restriction to a parametric family were relaxed,

$$-\log \frac{\hat{g}_{\text{alt}}(T(x_i))}{g_{\theta_0}(T(x_i))} = -\log \frac{1 - \widehat{\text{LFDR}}(x_i)}{\widehat{\text{LFDR}}(x_i)} + \log \frac{1-\hat{\pi}_0}{\hat{\pi}_0} \tag{20}$$

would be the information for discrimination according to the empirical Bayes methodology of Section 1.

The *regret* (Grünwald, 2007) of the codelength function $L_i^\dagger$ is

$$\operatorname{reg}\left(g_i^\dagger, x_i\right) = L_i^\dagger\left(T\left(x_i\right)\right) - \inf_{\theta\in\Theta}\left(-\log g_\theta\left(T\left(x_i\right)\right)\right) = -\log\frac{g_i^\dagger\left(T\left(x_i\right)\right)}{g_{\hat{\theta}}\left(T\left(x_i\right)\right)},$$

where $L_i^\dagger$ is given by $L_i^\dagger\left(T\left(x_i\right)\right) = -\log g_i^\dagger\left(T\left(x_i\right)\right)$ and where $\hat{\theta} = \arg\sup_{\theta\in\Theta} g_\theta\left(T\left(x\right)\right)$. Likewise, the regret of the codelength function relative to the null hypothesis is $\operatorname{reg}\left(g_{\theta_0}, x_i\right) = -\log\left(g_{\theta_0}\left(T\left(x_i\right)\right)/g_{\hat{\theta}}\left(T\left(x_i\right)\right)\right)$.

While the sign of $\Delta_i^\dagger\left(T\left(x_i\right)\right)$ indicates which hypothesis is favored (Rissanen, 1987), it can also be compared to a threshold J of the minimum amount of information considered sufficient for selecting one hypothesis over the other. In that case, the probability of observing misleading information for discrimination has an upper bound for any distributions g_{θ_0} and $g_i^\dagger$. Specifically, for any $J > 0$,

$$P_{\theta_0,\lambda}\left(\Delta_i^\dagger\left(T\left(X_i\right)\right) \leq -J\right) = P_{\theta_0,\lambda}\left(g_i^\dagger\left(T\left(X_i\right)\right)/g_{\theta_0}\left(T\left(X_i\right)\right) \geq \log^{-1} J\right) \leq 1/\log^{-1} J. \quad (21)$$

Applications to the probability of observing misleading evidence appear in Royall (2000). A derivation from the Markov inequality appears in Bickel (2012). Since the derivation assumes that g_{θ_0} and $g_i^\dagger$ are genuine probability density functions, formula (21) does not necessarily hold for pseudo-likelihoods such as profile likelihoods and likelihoods integrated with respect to an improper prior; however, it does hold for all marginal and conditional likelihoods (Royall, 2000).

The following two schemes († and ‡) for coding the reduced data give essentially identical regrets for a sufficiently large value of N.

Exact codelength

While the codelength function $L_i^\dagger$ for the ith feature cannot depend on x_i, it may depend on x_j for all $j \neq i$ as follows. For all $i = 1, \ldots, N$, define $L_i^\dagger$ such that the corresponding probability density function $g_i^\dagger$ is equal to $g_{\theta_i^\dagger}$ for the value $\theta_i^\dagger$ such that

$$\theta_i^\dagger = \arg\inf_{\theta \in \Theta} \sum_{j \neq i} \min\left(\operatorname{reg}\left(g_\theta, x_j\right), \operatorname{reg}\left(g_{\theta_0}, x_j\right)\right). \tag{22}$$

In words, the code for a given feature uses the distribution in the parametric family that minimizes the regret summed over all other features.

Proportional to N^2, the computation time can prohibit the use of the universal compression method for large N. For example, N can be in the tens of thousands for gene expression microarrays or in the hundreds of thousands for genome-wide association studies. The next coding scheme overcomes this problem because its computation time is proportional to N.

Approximate codelength

The † coding scheme is efficiently approximated by a slightly illegal scheme denoted by ‡. It determines the codelength for statistic $T(x_i)$ under the alternative hypothesis by using a common probability density function $g^\ddagger$ that is in the parametric family, i.e., $g^\ddagger = g_{\theta^\ddagger}$ for some $\theta^\ddagger \in \Theta$. This is accomplished by minimizing the regret over all features

$$\theta^\ddagger = \arg\inf_{\theta \in \Theta} \sum_{j=1}^{N} \min\left(\operatorname{reg}\left(g_\theta, x_j\right), \operatorname{reg}\left(g_{\theta_0}, x_j\right)\right). \tag{23}$$

This coding scheme is technically illegal in the sense that $g^\ddagger$, as a function of the observed data for each feature, depends on hindsight. However, under general conditions, $\theta^\ddagger$ approximates $\theta_i^\dagger$ for all $i = 1, \ldots, N$ given sufficiently large N because the selection of the distribution depends on all features without giving undue weight to any single feature. The approximation is supported by the fact that both $\theta^\dagger$ and $\theta^\ddagger$ are maximum likelihood estimates of θ under

the alternative hypothesis:

Theorem 1. *Assume that for some $\theta_0 \in \Theta$ and $\theta_{alt} \in \Theta$ such that $\theta_0 \neq \theta_{alt}$ and that for all $j \in \{1, \ldots, N\}$, each statistic $T(X_j)$ has probability density g_{θ_j} with $\theta_j \in \{\theta_0, \theta_{alt}\}$ and is independent of every $T(X_k)$ with $k \in \{1, \ldots, N\} \setminus \{j\}$. It follows that $\theta^{\ddagger}$, if unique, is the maximum likelihood estimate of θ_{alt}.*

Proof. Using equation (23),

$$\begin{aligned}
\theta^{\ddagger} &= \arg\inf_{\theta} \sum_{j=1}^{N} \min\left(-\log g_{\theta}\left(T\left(x_j\right)\right),\ -\log g_{\theta_0}\left(T\left(x_j\right)\right)\right) \\
&= \arg\sup_{\theta \in \Theta} \sum_{j=1}^{N} \max\left(\log g_{\theta}\left(T\left(x_j\right)\right),\ \log g_{\theta_0}\left(T\left(x_j\right)\right)\right) \\
&= \arg\sup_{\theta \in \Theta} \sup_{\boldsymbol{\theta} \in \{\theta_0, \theta_{\mathrm{alt}}\}^N} \sum_{j=1}^{N} \log g_{\theta_j}\left(T\left(x_j\right)\right) \\
&= \arg\sup_{\theta \in \Theta} \sup_{\boldsymbol{\theta} \in \{\theta_0, \theta_{\mathrm{alt}}\}^N} \prod_{j=1}^{N} g_{\theta_j}\left(T\left(x_j\right)\right),
\end{aligned}$$

where $\boldsymbol{\theta} = \langle \theta_1, \ldots, \theta_N \rangle$ and $\{\theta_0, \theta_{\mathrm{alt}}\}^N$ is the N-factor Cartesian product $\{\theta_0, \theta_{\mathrm{alt}}\} \times \cdots \times \{\theta_0, \theta_{\mathrm{alt}}\}$. □

Corollary 1. *Under the assumptions of Theorem 1, $i \in \{1, \ldots, N\}$, if $\theta_i^{\dagger}$ is unique, then it is the maximum likelihood estimate of θ_{alt} on the basis of the outcomes $X_j = x_j$ for all $j \in \{1, \ldots, N\} \setminus \{i\}$.*

Proof. The claim reduces to that of Theorem 1 because the data are equivalent except for the presence or absence of the outcome $T(X_i) = T(x_i)$ and because $\theta_i^{\dagger}$ and $\theta^{\ddagger}$ are equivalent, except for the presence or absence of the term involving that outcome. Thus, for all $i \in \{1, \ldots, N\}$,

$$\theta_i^{\dagger} = \arg\sup_{\theta \in \Theta} \sup_{\boldsymbol{\theta} \in \{\theta_0, \theta_{\mathrm{alt}}\}^N} \prod_{j \neq i} g_{\theta_j}\left(T\left(x_j\right)\right).$$

□

Theorem 3 of the next subsection specifies sufficient conditions for the convergence of $\theta^{\ddagger} - \theta_i^{\dagger}$ to 0 as N increases.

The coding method of the section entitled "Exact codelength" is *universal* in the sense that it asymptotically compresses the data as much as the noiseless coding theorem allows for any distribution in the parametric family (cf. Rissanen (2007, §3.7) and Grünwald (2007, §6.5)). Sufficient conditions for universality are stated in the following lemma, in which *strong consistency* means almost sure convergence to a parameter value as $n \to \infty$ if each $T(X_i)$ is stationary and, at fixed n, of a density function in $\{g_\theta : \theta \in \Theta\}$. (The dependence of g_θ on n is suppressed.) Such convergence will be denoted by $\xrightarrow{n}$.

Lemma 1 (Consistency)**.** *Suppose that for some $\theta_0 \in \Theta$ and $\theta_{\text{alt}} \in \Theta$ such that $\theta_0 \neq \theta_{\text{alt}}$ and that for all $j \in \{1, \ldots, N\}$, each statistic $T(X_j)$ has probability density g_{θ_j} with $\theta_j \in \{\theta_0, \theta_{\text{alt}}\}$ such that $\theta_j = \theta_{\text{alt}}$ for at least two values of j in $\{1, \ldots, N\}$. Suppose further that $g_\bullet(T(X_j))$ is almost surely continuous on Θ for all $j \in \{1, \ldots, N\}$. If, for some $i \in \{1, \ldots, N\}$, $\theta_i^{\dagger}$ is unique and $\hat{\theta}_j = \arg\sup_{\theta \in \Theta} g_\theta(T(X_j))$ is a strongly consistent estimate of θ_j for all $j \in \{1, \ldots, N\} \setminus \{i\}$, then $\theta_i^{\dagger}$ is a strongly consistent estimate of θ_{alt}.*

Proof. Let $\mathfrak{J} = \{j : \theta_j = \theta_{\text{alt}}, j \in \{1, \ldots, N\} \setminus \{i\}\}$, which by assumption is nonempty. By the consistency condition, $\hat{\theta}_j \xrightarrow{n} \theta_{\text{alt}}$ for all $j \in \mathfrak{J}$ and $\hat{\theta}_j \xrightarrow{n} \theta_0$ for all $j \in \{1, \ldots, N\} \setminus \mathfrak{J}$. Thus, with probability 1,

$$\begin{aligned}
\prod_{j \neq i} g_{\theta_j}(T(X_j)) &= \prod_{j \in \mathfrak{J}} g_{\theta_{\text{alt}}}(T(X_j)) \prod_{j \notin \mathfrak{J} \cup \{i\}} g_{\theta_0}(T(X_j)) \\
&= \prod_{j \neq i} g_{\hat{\theta}_j}(T(X_j)) \\
&= \prod_{j \neq i} \max\left(g_{\theta_{\text{alt}}}(T(X_j)), g_{\theta_0}(T(X_j))\right) \\
&= \sup_{\theta \in \Theta} \prod_{j \neq i} \max\left(g_\theta(T(X_j)), g_{\theta_0}(T(X_j))\right)
\end{aligned}$$

in the limit as $n \to \infty$, with the equalities holding by the almost-sure continuity of $g_\bullet(T(X_j))$

as a function on Θ (Serfling, 1980, §1.7). Because by equation (22),

$$\theta_i^\dagger = \arg\sup_{\theta\in\Theta} \sum_{j\neq i} \max\left(g_\theta\left(T\left(X_j\right)\right), g_{\theta_0}\left(T\left(X_j\right)\right)\right),$$

it follows that $\theta_i^\dagger \overset{n}{\to} \theta_i$. □

Heuristically, the key observation of the proof is that whether θ is constrained to have one of the two values has no asymptotic effect on the estimates of θ_j. The universality of the codelength function is a consequence.

Theorem 2 (Universality)**.** *Under the conditions of Lemma 1,*

$$\lim_{n\to\infty} E_{\theta_{\text{alt}}}\left(\frac{L_i^\dagger\left(T\left(X_i\right)\right)}{n}\right) = \lim_{n\to\infty} E_{\theta_{\text{alt}}}\left(\frac{-\log g_{\theta_{alt}}\left(T\left(X_i\right)\right)}{n}\right)$$

for all $i \in \{1,\ldots,N\}$ such that $\theta_i = \theta_{\text{alt}}$, where $E_{\theta_{\text{alt}}}$ signifies the expectation value with respect to $g_{\theta_{\text{alt}}}$, i.e., $E_{\theta_{\text{alt}}}(\bullet) = \int \bullet dP_{\theta_{\text{alt}}}$.

Proof. $P_{\theta_{\text{alt}}}\left(\lim_{n\to\infty}\theta_i^\dagger \in \{\theta_0,\theta_{\text{alt}}\}\right) = 1$ for all $i \in \{1,\ldots,N\}$ because $\theta_i^\dagger \overset{n}{\to} \theta_i$ by the lemma and $\theta_i \in \{\theta_0,\theta_{\text{alt}}\}$ by assumption. Hence, $\theta_i^\dagger \overset{n}{\to} \theta_{\text{alt}}$ for all $i \in \{1,\ldots,N\}$ such that $\theta_i = \theta_{\text{alt}}$. Thus, for those values of i,

$$\lim_{n\to\infty} E_{\theta_{\text{alt}}}\left(\frac{-\log\left(g_{\theta_{\text{alt}}}\left(T\left(X_i\right)\right)/g_{\theta_i^\dagger}\left(T\left(X_i\right)\right)\right)}{n}\right) = 0$$

because $g_{\theta_{\text{alt}}}\left(T\left(X_i\right)\right)/g_{\theta_i^\dagger}\left(T\left(X_i\right)\right) \overset{n}{\to} 1$ by the almost-sure continuity of $g_\bullet\left(T\left(X_i\right)\right)$ as a function on Θ (Serfling, 1980, §1.7). □

The $N \to \infty$ universally of a related mixture code will be established in Appendix B.

Asymptotic characteristics of $\theta^{\ddagger}$ and $\theta_i^{\dagger}$

Assume $X_1, X_2, \ldots$ are independent and each of identical distribution $P_\star$. For example, $P_\star$ could be a K-component mixture distribution $P_\star = \sum_{k=1}^{K} \pi_k P_{\star k}$, where π_k is the probability that some X_j has distribution $P_{\star k}$, which is not necessarily in $\{P_{\theta,\lambda} : \theta \in \Theta, \lambda \in \Lambda\}$. Let $E_\star(\bullet)$ and $\overset{N}{\to}$ denote the expectation value and almost-sure convergence as $N \to \infty$ with respect to $P_\star$.

Theorem 3. Suppose that, for all $i \in \{1, \ldots, N\}$, $E_\star(\log g_\theta(T(X_j))) < \infty$ for all $\theta \in \Theta$ and that $\theta^{\ddagger}$ and $\theta_i^{\dagger}$ are unique with $P_\star$-probability 1. Then $\theta^{\ddagger} - \theta_i^{\dagger} \overset{N}{\to} 0$ *for all* $i \in \{1, \ldots, N\}$.

Proof. For any $\theta \in \Theta$, let $\hat{\theta}_j(\theta) = \arg\max_{\tilde{\theta} \in \{\theta_0, \theta\}} g_{\tilde{\theta}}(T(x_j))$. Because $\log g_{\hat{\theta}_j(\theta)}(T(X_j))$ is IID for all $j \in \{1, \ldots, N\}$, the strong law of large numbers implies that, for all $\mathcal{J}_N \in \{\{1, \ldots, N\}, \{1, \ldots, N\} \setminus \{1\}, \ldots, \{1, \ldots, N\} \setminus \{N\}\}$,

$$\frac{1}{|\mathcal{J}_N|} \sum_{j \in \mathcal{J}_N} \log g_{\hat{\theta}_j(\theta)}(T(X_j)) \overset{N}{\to} E_\star\left(\log g_{\hat{\theta}_j(\theta)}(T(X_j))\right)$$

$$= P_\star\left(\hat{\theta}_j(\theta) = \theta_0\right) E_\star\left(\log g_{\hat{\theta}_j(\theta)}(T(X_j)) \,|\hat{\theta}_j(\theta) = \theta_0\right)$$

$$+ P_\star\left(\hat{\theta}_j(\theta) = \theta\right) E_\star\left(\log g_{\hat{\theta}_j(\theta)}(T(X_j)) \,|\hat{\theta}_j(\theta) = \theta\right),$$

the finiteness of which follows from that of $E_\star(\log g_\theta(T(X_j)))$. As the result holds for arbitrary $\theta \in \Theta$,

$$\arg\sup_{\theta \in \Theta} \frac{1}{|\mathcal{J}_N|} \sum_{j \in \mathcal{J}_N} \log g_{\hat{\theta}_j(\theta)}(T(X_j)) \overset{N}{\to}$$

$$\arg\sup_{\theta \in \Theta} E_\star\left(\log g_{\hat{\theta}_i(\theta)}(T(X_j))\right)$$

irrespective of whether the sum on the left-hand-side is over $\{1, \ldots, N\}$ or over $\{1, \ldots, N\} \setminus \{i\}$ for some $i \in \{1, \ldots, N\}$. (The uniqueness of the maximizing value of θ on the left-hand-side is guaranteed by the postulated uniqueness of $\theta^{\ddagger}$ and $\theta_i^{\dagger}$.) Therefore, the difference in the maximum likelihood estimate of θ under the alternative hypothesis using $X_1, \ldots, X_N$ and

that using $X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_N$ converges almost surely to 0; however, such maximum likelihood estimates are $\theta^{\ddagger}$ and $\theta_i^{\dagger}$, respectively, according to Theorem 1 and Corollary 1. $\square$

Appendix B: Asymptotic characteristics of MDL and L0O

This section extends the fixed-component results of Appendix A to the two-component mixture density of equation (3) with the constraint that $g_{\text{alt}} = g_{\theta_{\text{alt}}}$ for some $\theta_{\text{alt}} \in \Theta$. In this setting, the universal density $g_i^{\dagger}$ and its approximation $g^{\ddagger}$ are replaced with $g_i^{\text{MDL}} = g_{\theta_i^{\text{MDL}}}$ and its approximation $g^{\text{L0O}} = g_{\theta^{\text{L0O}}}$, where $\left\langle \theta_i^{\text{MDL}}, \pi_{0i}^{\text{MDL}} \right\rangle$ are given by equation (12). (Yang and Bickel (2010) compared the performance of $g^{\ddagger}$ and g^{L0O} by simulation.)

Assuming the statistics are independent, $\left\langle \theta_i^{\text{MDL}}, \pi_{0i}^{\text{MDL}} \right\rangle$ and $\left\langle \theta^{\text{L0O}}, \pi_0^{\text{L0O}} \right\rangle$ are clearly maximum likelihood estimates of $\langle \theta_{\text{alt}}, \pi_0 \rangle$. Consequently, the steps used to prove Theorem 3 also demonstrate that $\theta^{\text{L0O}} - \theta_i^{\text{MDL}} \xrightarrow{N} 0$ and $\pi_0^{\text{L0O}} - \pi_{0i}^{\text{MDL}} \xrightarrow{N} 0$ for all $i \in \{1, \ldots, N\}$ under the independence condition. The mixture codes form LFDR estimates via substituting either θ^{MDL} and π_0^{MDL} or θ^{L0O} and π_0^{L0O} into equations (2) and (3).

Whereas regularity conditions entailing the strong consistency of maximum likelihood estimates for finite-mixture models (Redner and Walker, 1984) would apply as $N \to \infty$, seemingly more pertinent to universality is consistency in the sense of $\xrightarrow{n}$, which is almost-sure convergence as $n \to \infty$ under the stationarity of every $T(X_i)$. However, such $\xrightarrow{n}$ consistency does not hold if N is finite and if $\pi_0 > 0$, for in that case, there is fixed, nonzero probability π_0^N that all N statistics have probability density function g_{θ_0} rather than $g_{\theta_{\text{alt}}}$. Therefore, $\xrightarrow{N}$ consistency will be used instead.

Theorem 4. *If the maximum likelihood estimate θ^{L0O} almost surely converges to θ_{alt} as $N \to \infty$ and if $g_{\bullet}(T(X_i))$ is almost surely continuous on Θ for all $i \in \{1, 2, \ldots\}$, then*

$$\lim_{N \to \infty} E_{\theta_{\text{alt}}} \left(L_i^*(T(X_i)) / n \right) = E_{\theta_{\text{alt}}} \left(-\log g_{\theta_{\text{alt}}}(T(X_i)) / n \right)$$

for all $i \in \{1, 2, \ldots\}$ such that $\theta_i = \theta_{\text{alt}}$, where $L_i^*(T(X_i)) = -\log g_{\theta_i^{\text{MDL}}}(T(X_j))$ and $E_{\theta_{\text{alt}}}$

signifies the expectation value with respect to $g_{\theta_{\rm alt}}$*, i.e.,* $E_{\theta_{\rm alt}}(\bullet) = \int \bullet dP_{\theta_{\rm alt}}$.

Proof. Since $\theta_i^{\rm MDL}$ is the maximum likelihood estimate for the $N-1$ statistics other than $T(X_i)$, $\theta_i^{\rm MDL} \overset{N}{\to} \theta_{\rm alt}$. Thus, the claim follows from reasoning analogous to that used to prove Theorem 2. □

Corollary 2 (Asymptotic universality)**.** *Given the conditions of Theorem 4,*

$$\lim_{n\to\infty}\lim_{N\to\infty} E_{\theta_{\rm alt}}\left(\frac{L_i^*\left(T\left(X_i\right)\right)}{n}\right)$$

$$= \lim_{n\to\infty} E_{\theta_{\rm alt}}\left(\frac{-\log g_{\theta_{\rm alt}}\left(T\left(X_j\right)\right)}{n}\right)$$

for all $i \in \{1, 2, \dots\}$ *such that* $\theta_i = \theta_{\rm alt}$.

The proof is trivial. The corollary means that

$$\left(L_i^*\left(T\left(x_i\right)\right) - \log\left(1-\pi_{0i}^{\rm MDL}\right)\right) - \left(L^0\left(T\left(x_i\right)\right) - \log\pi_{0i}^{\rm MDL}\right)$$

may be regarded as approaching the information for discrimination under the mixture model as $N\to\infty$. Since $\theta^{\rm L0O} - \theta_i^{\rm MDL} \overset{N}{\to} 0$ and $\pi_0^{\rm L0O} - \pi_{0i}^{\rm MDL} \overset{N}{\to} 0$, that information is approximated by substituting the maximum likelihood estimates $\theta^{\rm L0O}$ and $\pi_0^{\rm L0O}$ for $\theta_i^{\rm MDL}$ and $\pi_{0i}^{\rm MDL}$, respectively.

References

Aitkin, M., 1991. Posterior Bayes factors (with discussion). Journal of the Royal Statistical Society B 53, 111–142.

Barron, A., Rissanen, J., Yu, B., 1998. The minimum description length principle in coding and modeling. IEEE Transactions on Information Theory 44, 2743–2760.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society B 57, 289–300.

Bickel, D. R., 2002. Microarray gene expression analysis: Data transformation and multiple-comparison bootstrapping. Computing Science and Statistics 34, 383–400.

Bickel, D. R., 2011a. Measuring support for a hypothesis about a random parameter without estimating its unknown prior. Technical Report, Ottawa Institute of Systems Biology, arXiv:1101.0305.

Bickel, D. R., 2011b. A predictive approach to measuring the strength of statistical evidence for single and multiple comparisons. Canadian Journal of Statistics 39, 610–631.

Bickel, D. R., 2011c. Resolving conflicts between statistical methods by probability combination: Application to empirical Bayes analyses of genomic data. Technical Report, Ottawa Institute of Systems Biology, arXiv:1111.6174.

Bickel, D. R., 2011d. Simple estimators of false discovery rates given as few as one or two p-values without strong parametric assumptions. Technical Report, Ottawa Institute of Systems Biology, arXiv:1106.4490.

Bickel, D. R., 2011e. Small-scale inference: Empirical Bayes and confidence methods for as few as a single comparison. Technical Report, Ottawa Institute of Systems Biology, arXiv:1104.0341.

Bickel, D. R., 2012. Empirical Bayes interval estimates that are conditionally equal to unadjusted confidence intervals or to default prior credibility intervals. To appear in Statistical Applications in Genetics and Molecular Biology; 2010 version available from arXiv:1012.6033.

Dabney, A., Storey, J. D., with assistance from Gregory R. Warnes, 2011. qvalue: Q-value estimation for false discovery rate control. Reference Manual, R package version 1.26.0.

Efron, B., 2004. Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. Journal of the American Statistical Association 99, 96–104.

Efron, B., 2007. Size, power and false discovery rates. Annals of Statistics 35, 1351–1377.

Efron, B., 2008. Simultaneous inference: When should hypothesis testing problems be combined? Ann. Appl. Statist 2, 197–223.

Efron, B., 2010. Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction. Cambridge University Press, Cambridge.

Efron, B., Tibshirani, R., 2002. Empirical Bayes methods and false discovery rates for microarrays. Genetic Epidemiology 23, 70–86.

Efron, B., Tibshirani, R., Storey, J. D., Tusher, V., 2001. Empirical Bayes analysis of a microarray experiment. J. Am. Stat. Assoc. 96, 1151–1160.

Gastpar, M., Gill, P., Huth, A., Theunissen, F., 2010. Anthropic correction of information estimates and its application to neural coding. IEEE Transactions on Information Theory 56, 890–900.

Gentleman, R. C., Carey, V. J., Bates, D. M., et al., 2004. Bioconductor: Open software development for computational biology and bioinformatics. Genome Biology 5, R80.

Goodman, D., 2004. Taking the prior seriously: Bayesian analysis without subjective probability. The Nature of Scientific Evidence: Statistical, Philosophical, and Empirical Considerations. University of Chicago Press, Chicago, pp. 379–400.

Grünwald, P. D., 2007. The Minimum Description Length Principle. MIT Press, London.

Hong, W.-J., Tibshirani, R., Chu, G., 2009. Local false discovery rate facilitates comparison of different microarray experiments. Nucleic Acids Research 37 (22), 7483–7497.

Hu, F. F., Zidek, J. V., 2002. The weighted likelihood. Canadian Journal of Statistics 30, 347–371.

Li, X., 2009. ProData. Bioconductor.org documentation for the ProData package.

Montazeri, Z., Yanofsky, C. M., Bickel, D. R., 2010. Shrinkage estimation of effect sizes as an alternative to hypothesis testing followed by estimation in high-dimensional biology: Applications to differential gene expression. Statistical Applications in Genetics and Molecular Biology 9, 23.

Muralidharan, O., 2010. An empirical Bayes mixture method for effect size and false discovery rate estimation. Annals of Applied Statistics 4, 422–438.

Pawitan, Y., 2001. In All Likelihood: Statistical Modeling and Inference Using Likelihood. Clarendon Press, Oxford.

R Development Core Team, 2008. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Redner, R. A., Walker, H. F., 1984. Mixture densities, maximum likelihood and the EM algorithm. SIAM Review 26, 195–239.

Rissanen, J., 1987. Stochastic complexity. Journal of the Royal Statistical Society B 49, 223–239.

Rissanen, J., 2007. Information and Complexity in Statistical Modeling. Springer, New York.

Royall, R., 2000. On the probability of observing misleading statistical evidence. Journal of the American Statistical Association 95, 760–768.

Ruckdeschel, P., Kohl, M., Stabla, T., Camphausen, F., May 2006. S4 classes for distributions. R News 6 (2), 2–6.

Schweder, T., Hjort, N. L., 2002. Confidence and likelihood. Scandinavian Journal of Statistics 29, 309–332.

Seifert, E. L., Fiehn, O., Bezaire, V., Bickel, D. R., Wohlgemuth, G., Adams, S. H., Harper, M.-E., 2010. Long-chain fatty acid combustion rate is associated with unique metabolite profiles in skeletal muscle mitochondria. PLoS ONE 5, e9834.

Serfling, R. J., 1980. Approximation theorems of mathematical statistics. Wiley, New York.

Severini, T., 2000. Likelihood Methods in Statistics. Oxford University Press, Oxford.

Storey, J. D., 2002. A direct approach to false discovery rates. Journal of the Royal Statistical Society. Series B: Statistical Methodology 64, 479–498.

Wuertz, D., 2010. fbasics: Rmetrics - markets and basic statistics. Reference Manual, R package version 2110.79.

Yang, Y., Bickel, D. R., 2010. Minimum description length and empirical Bayes methods of identifying SNPs associated with disease. Technical Report, Ottawa Institute of Systems Biology, COBRA Preprint Series, Article 74, biostats.bepress.com/cobra/ps/art74.

Yang, Z., Li, Z., Bickel, D. R., 2011. Empirical Bayes estimation of posterior probabilities of enrichment. Technical Report, Ottawa Institute of Systems Biology, arXiv:1201.0153.

Yanofsky, C. M., Bickel, D. R., 2010. Validation of differential gene expression algorithms: Application comparing fold-change estimation to hypothesis testing. BMC Bioinformatics 11, 63.